Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization

Shai Shalev-Shwartz
School of Computer Science and Engineering
Hebrew University, Jerusalem, Israel

Tong Zhang
Department of Statistics
Rutgers University, NJ, USA

Abstract

Stochastic Gradient Descent (SGD) has become popular for solving large scale supervised machine learning optimization problems such as SVM, due to their strong theoretical guarantees. While the closely related Dual Coordinate Ascent (DCA) method has been implemented in various software packages, it has so far lacked good convergence analysis. This paper presents a new analysis of Stochastic Dual Coordinate Ascent (SDCA) showing that this class of methods enjoy strong theoretical guarantees that are comparable or better than SGD. This analysis justifies the effectiveness of SDCA for practical applications.

1 Introduction

We consider the following generic optimization problem associated with regularized loss minimization of linear predictors: Let $x_1, \ldots, x_n$ be vectors in $\mathbb{R}^d$, let $\phi_1, \ldots, \phi_n$ be a sequence of scalar convex functions, and let $\lambda > 0$ be a regularization parameter. Our goal is to solve $\min_{w \in \mathbb{R}^d} P(w)$ where

$$P(w) = \left[\frac{1}{n} \sum_{i=1}^{n} \phi_i(w^\top x_i) + \frac{\lambda}{2} \|w\|^2 \right]. \tag{1}$$

For example, given labels $y_1, \ldots, y_n$ in $\{\pm 1\}$, the SVM problem (with linear kernels and no bias term) is obtained by setting $\phi_i(a) = \max\{0, 1 - y_i a\}$. Regularized logistic regression is obtained by setting $\phi_i(a) = \log(1 + \exp(-y_i a))$. Regression problems also fall into the above. For example, ridge regression is obtained by setting $\phi_i(a) = (a - y_i)^2$, regression with the absolute-value is obtained by setting $\phi_i(a) = |a - y_i|$, and support vector regression is obtained by setting $\phi_i(a) = \max\{0, |a - y_i| - \nu\}$, for some predefined insensitivity parameter $\nu > 0$.

Let w^* be the optimum of (1). We say that a solution w is ϵ_P-sub-optimal if $P(w) - P(w^*) \leq \epsilon_P$. We analyze the runtime of optimization procedures as a function of the time required to find an ϵ_P-sub-optimal solution.

A simple approach for solving SVM is stochastic gradient descent (SGD) [13, 11, 1]. SGD finds an ϵ_P-sub-optimal solution in time $\tilde{O}(1/(\lambda \epsilon_P))$. This runtime does not depend on n and therefore is favorable when n is very large. However, the SGD approach has several disadvantages. It does not have a clear stopping criterion; It tends to be too aggressive at the beginning of the optimization process, especially when λ is very small; While SGD reaches a moderate accuracy quite fast, it's convergence becomes rather slow when we are interested in more accurate solutions.

An alternative approach is dual coordinate ascent (DCA), which solves a *dual* problem of (1). Specifically, for each i let $\phi_i^* : \mathbb{R} \to \mathbb{R}$ be the convex conjugate of ϕ_i, namely, $\phi_i^*(u) = \max_z(zu - \phi_i(z))$. The dual problem is

$$\max_{\alpha \in \mathbb{R}^m} D(\alpha) \quad \text{where} \quad D(\alpha) = \left[\frac{1}{n} \sum_{i=1}^n -\phi_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i x_i \right\|^2 \right]. \tag{2}$$

The dual objective in (2) has a different dual variable associated with each example in the training set. At each iteration of DCA, the dual objective is optimized with respect to a single dual variable, while the rest of the dual variables are kept in tact.

If we define

$$w(\alpha) = \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i x_i, \tag{3}$$

then it is known that $w(\alpha^*) = w^*$, where α^* is an optimal solution of (2). It is also known that $P(w^*) = D(\alpha^*)$ which immediately implies that for all w and α, we have $P(w) \geq D(\alpha)$, and hence the duality gap defined as

$$P(w(\alpha)) - D(\alpha)$$

can be regarded as an upper bound of the primal sub-optimality $P(w(\alpha)) - P(w^*)$.

We focus on a *stochastic* version of DCA, abbreviated by SDCA, in which at each round we choose which dual coordinate to optimize uniformly at random. The purpose of this paper is to develop theoretical understanding of the convergence of the duality gap for SDCA.

We analyze SDCA either for L-Lipschitz loss functions or for $(1/\gamma)$-smooth loss functions, which are defined as follows.

Definition 1. *A function $\phi_i : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz if for all $a, b \in \mathbb{R}$, we have*

$$|\phi_i(a) - \phi_i(b)| \leq L |a - b|.$$

A function $\phi_i : \mathbb{R} \to \mathbb{R}$ is $(1/\gamma)$-smooth if it is differentiable and its derivative is $(1/\gamma)$-Lipschitz. An equivalent condition is that for all $a, b \in \mathbb{R}$, we have

$$\phi_i(a) \leq \phi_i(b) + \phi_i'(b)(a - b) + \frac{1}{2\gamma}(a - b)^2.$$

It is well-known that if $\phi_i(a)$ is $(1/\gamma)$-smooth, then $\phi_i^*(u)$ is γ strongly convex: for all $u, v \in \mathbb{R}$ and $s \in [0, 1]$:

$$-\phi_i^*(su + (1 - s)v) \geq -s\phi_i^*(u) - (1 - s)\phi_i^*(v) + \frac{\gamma s(1 - s)}{2}(u - v)^2.$$

Our main findings are: in order to achieve a duality gap of ϵ,

- For L-Lipschitz loss functions, we obtain the rate of $\tilde{O}(n + L^2/(\lambda\epsilon))$.

- For $(1/\gamma)$-smooth loss functions, we obtain the rate of $\tilde{O}((n + 1/(\lambda\gamma))\log(1/\epsilon))$.

- For loss functions which are almost everywhere smooth (such as the hinge-loss), we can obtain rate better than the above rate for Lipschitz loss. See Section 5 for a precise statement.

2

2 Related Work

DCA methods are related to decomposition methods [10, 5]. While several experiments have shown that decomposition methods are inferior to SGD for large scale SVM [11, 6], Hsieh et al. [3] recently argued that SDCA outperform the SGD approach in some regimes. For example, this occurs when we need relatively high solution accuracy so that either SGD or SDCA has to be run for more than a few passes over the data.

However, our theoretical understanding of SDCA is not satisfying. Several authors (e.g. [8, 3]) proved a linear convergence rate for solving SVM with DCA (not necessarily stochastic). The basic technique is to adapt the linear convergence of coordinate ascent that was established by Luo and Tseng [7]. The linear convergence means that it achieves a rate of $(1 - \nu)^k$ after k passes over the data, where $\nu > 0$. This convergence result tells us that after an unspecified number of iterations, the algorithm converges faster to the optimal solution than SGD.

However, there are two problems with this analysis. First, the linear convergence parameter, ν, may be very close to zero and the initial unspecified number of iterations might be very large. In fact, while the result of [7] does not explicitly specify ν, an examine of their proof shows that ν is proportional to the smallest nonzero eigenvalue of $X^\top X$, where X is the $n \times d$ data matrix with its i-th row be the i-th data point x_i. For example if two data points $x_i \neq x_j$ becomes closer and closer, then $\nu \to 0$. This dependency is problematic in the data laden domain, and we note that such a dependency does not occur in the analysis of SGD.

Second, the analysis only deals with the sub-optimality of the *dual* objective, while our real goal is to bound the sub-optimality of the *primal* objective. Given a dual solution $\alpha \in \mathbb{R}^n$ its corresponding primal solution is $w(\alpha)$ (see (3)). The problem is that even if α is ϵ_D-sub-optimal in the dual, for some small ϵ_D, the primal solution $w(\alpha)$ might be far from being optimal. For SVM, [4, Theorem 2] showed that in order to obtain a primal ϵ_P-sub-optimal solution, we need a dual ϵ_D-sub-optimal solution with $\epsilon_D = O(\lambda \epsilon_P^2)$; therefore a convergence result for dual solution can only translate into a primal convergence result with worse convergence rate. Such a treatment is unsatisfactory, and this is what we will avoid in the current paper.

Some analyses of stochastic coordinate ascent provide solutions to the first problem mentioned above. For example, Collins et al [2] analyzed an exponentiated gradient dual coordinate ascent algorithm. The algorithm analyzed there (exponentiated gradient) is different from the standard DCA algorithm which we consider here, and the proof techniques are quite different. Consequently their results are not directly comparable to results we obtain in this paper. Nevertheless we note that for SVM, their analysis shows a convergence rate of $O(n/\epsilon_D)$ in order to achieve ϵ_D-sub-optimality (on the dual) while our analysis shows a convergence of $O(n \log \log n + 1/\lambda \epsilon)$ to achieve ϵ duality gap; for logistic regression, their analysis shows a convergence rate of $O(n^2 \log(1/\epsilon_D))$ in order to achieve ϵ_D-sub-optimality on the dual while our analysis shows a convergence of $O((n + 1/\lambda) \log(1/\epsilon))$ to achieve ϵ duality gap.

In addition, [12], and later [9] have analyzed randomized versions of coordinate descent for unconstrained and constrained minimization of smooth convex functions. [3, Theorem 4] applied these results to the dual SVM formulation. However, the resulting convergence rate is $O(n/\epsilon_D)$ which is, as mentioned before, inferior to the results we obtain here. Furthermore, neither of these analyses can be applied to logistic regression due to their reliance on the smoothness of the dual objective function which is not satisfied for the dual formulation of logistic regression. We shall also point out again that all of these bounds are for the dual sub-optimality, while as mentioned before, we are interested in the primal sub-optimality.

In this paper we derive new bounds on the duality gap (hence, they also imply bounds on the primal sub-optimality) of SDCA. These bounds are superior to earlier results, and our analysis only holds for

randomized (stochastic) dual coordinate ascent. As we will see from our experiments, randomization is important in practice. In fact, the practical convergence behavior of (non-stochastic) cyclic dual coordinate ascent (even with a random ordering of the data) can be slower than our theoretical bounds for SDCA, and thus cyclic DCA is inferior to SDCA. In this regard, we note that some of the earlier analysis such as [7] can be applied both to stochastic and to cyclic dual coordinate ascent methods with similar results. This means that their analysis, which can be no better than the behavior of cyclic dual coordinate ascent, is inferior to our analysis.

The following table summarizes our results in comparison to previous analyses. Note that for SDCA with Lipschitz loss, we observe a faster practical convergence rate, which is explained with our refined analysis in Section 5.

<div align="center">

Lipschitz loss

Algorithm	type of convergence	rate
SGD	primal	$\tilde{O}(\frac{1}{\lambda\epsilon})$
online EG [2]	dual	$\tilde{O}(\frac{n}{\epsilon})$
SDCA	primal-dual	$\tilde{O}(n + \frac{1}{\lambda\epsilon})$ or faster

Smooth loss

Algorithm	type of convergence	rate
SGD	primal	$\tilde{O}(\frac{1}{\lambda\epsilon})$
online EG [2]	dual	$\tilde{O}(n^2 \log\frac{1}{\epsilon})$
SDCA	primal-dual	$\tilde{O}((n + \frac{1}{\lambda}) \log\frac{1}{\epsilon})$

</div>

3 Basic Results

The generic algorithm we analyze is as follows.

Procedure SDCA$(\alpha^{(0)})$

Let $w^{(0)} = w(\alpha^{(0)})$
Iterate: for $t = 1, 2, \ldots, T$:
  Randomly pick i
  Find $\Delta\alpha_i$ to maximize $-\phi_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i\|_2^2$.
  $\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$
  $w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i$
Output (Averaging option):
  Let $\bar{\alpha} = \frac{1}{T - T_0}\sum_{i=T_0+1}^{T} \alpha^{(t-1)}$
  Let $\bar{w} = w(\bar{\alpha}) = \frac{1}{T - T_0}\sum_{i=T_0+1}^{T} w^{(t-1)}$
  return $\bar{w}$
Output (Random option):
  Let $\bar{\alpha} = \alpha^{(t)}$ and $\bar{w} = w^{(t)}$ for some random $t \in T_0 + 1, \ldots, T$
  return $\bar{w}$

We analyze the algorithm based on different assumptions on the loss functions. To simplify the statements of our theorems, we always assume the following:

1. For all i, $\|x_i\| \leq 1$

2. For all i, $\phi_i(a) \geq 0$

3. For all i, $\phi_i(0) \leq 1$

Theorem 1. *Consider Procedure SDCA with $\alpha^{(0)} = 0$. Assume that ϕ_i is L-Lipschitz for all i. To obtain a duality gap of $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon_P$, it suffices to have a total number of iterations of*

$$T \geq T_0 + n + \frac{4\,L^2}{\lambda\epsilon_P} \geq \max(0, \lceil n\log(0.5\lambda nL^{-2})\rceil) + n + \frac{20\,L^2}{\lambda\epsilon_P}\ .$$

Moreover, when $t \geq T_0$, we have dual sub-optimality bound of $\mathbb{E}[D(\alpha^*) - D(\alpha^{(t)})] \leq \epsilon_P/2$.

Remark 1. *If we choose the average version, we may simply take $T = 2T_0$. Moreover, we note that Theorem 1 holds for both averaging or for choosing w at random from $\{T_0 + 1, \ldots, T\}$. This means that calculating the duality gap at few random points would lead to the same type of guarantee with high probability. This approach has the advantage over averaging, since it is easier to implement the stopping condition (we simply check the duality gap at some random stopping points. This is in contrast to averaging in which we need to know T, T_0 in advance).*

Remark 2. *The above theorem applies to the hinge-loss function, $\phi_i(u) = \max\{0, 1 - y_i a\}$. However, for the hinge-loss, the constant 4 in the first inequality can be replaced by 1 (this is because the domain of the dual variables is positive, hence the constant 4 in Lemma 4 can be replaced by 1). We therefore obtain the bound:*

$$T \geq T_0 + n + \frac{L^2}{\lambda\epsilon_P} \geq \max(0, \lceil n\log(0.5\lambda nL^{-2})\rceil) + n + \frac{5\,L^2}{\lambda\epsilon_P}\ .$$

Theorem 2. *Consider Procedure SDCA with $\alpha^{(0)} = 0$. Assume that ϕ_i is $(1/\gamma)$-smooth for all i. To obtain an expected duality gap of $\mathbb{E}[P(w^{(T)}) - D(\alpha^{(T)})] \leq \epsilon_P$, it suffices to have a total number of iterations of*

$$T \geq \left(n + \tfrac{1}{\lambda\gamma}\right)\log((n + \tfrac{1}{\lambda\gamma}) \cdot \tfrac{1}{\epsilon_P}).$$

Moreover, to obtain an expected duality gap of $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon_P$, it suffices to have a total number of iterations of

$$T_0 \geq \left(n + \tfrac{1}{\lambda\gamma}\right)\log((n + \tfrac{1}{\lambda\gamma}) \cdot \tfrac{1}{(T-T_0)\epsilon_P}).$$

Remark 3. *If we choose $T = 2T_0$, and assume that $T_0 \geq n + 1/(\lambda\gamma)$, then the second part of Theorem 2 implies a requirement of*

$$T_0 \geq \left(n + \tfrac{1}{\lambda\gamma}\right)\log(\tfrac{1}{\epsilon_P}),$$

which is slightly weaker than the first part of Theorem 2 when ϵ_P is relatively large.

Remark 4. *The estimation error of the primal objective behaves like $\Theta\left(\frac{1}{\lambda n}\right)$. Therefore, an interesting regime is when $\frac{1}{\lambda n} = \Theta(\epsilon)$. In that case, the bound for both Lipschitz and smooth functions would be $\tilde{O}(n)$. Another interesting regime is when we would like $\epsilon \ll \frac{1}{\lambda n}$, but still $\frac{1}{\lambda n} = O(1)$. In that case, smooth functions still yield the bound $\tilde{O}(n)$, but the dominating term for Lipschitz functions will be $\frac{1}{\lambda\epsilon}$.*

Remark 5. *The runtime of SGD is $\tilde{O}(\frac{1}{\lambda\epsilon})$. This can be better than SDCA if $n \gg \frac{1}{\lambda\epsilon}$. However, in that case, SGD in fact only looks at $n' = \tilde{O}(\frac{1}{\lambda\epsilon})$ examples, so we can run SDCA on these n' examples and obtain basically the same rate. For smooth functions, SGD can be much worse than SDCA if $\epsilon \ll \frac{1}{\lambda n}$.*

4 Using SGD at the first epoch

From the convergence analysis, SDCA may not perform as well as SGD for the first few epochs (each echo means one pass over the data). The main reason is that SGD takes a larger step size than SDCA earlier on, which helps its performance. It is thus natural to combine SGD and SDCA, where the first epoch is performed using a modified stochastic gradient descent rule. We show that the expected dual sub-optimality at the end of the first epoch is $\tilde{O}(1/(\lambda n))$. This result can be combined with SDCA to obtain a faster convergence.

We first introduce convenient notation. Let P_t denote the primal objective for the first t examples,

$$P_t(w) = \left[\frac{1}{t} \sum_{i=1}^{t} \phi_i(w^\top x_i) + \frac{\lambda}{2} \|w\|^2 \right].$$

The corresponding dual objective is

$$D_t(\alpha) = \left[\frac{1}{t} \sum_{i=1}^{t} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda t} \sum_{i=1}^{t} \alpha_i x_i \right\|^2 \right].$$

Note that $P_n(w)$ is the primal objective given in (1) and that $D_n(\alpha)$ is the dual objective given in (2).

The following algorithm is a modification of SGD. The idea is to greedily decrease the dual sub-optimality for problem $D_t(\cdot)$ at each step t. This is different from DCA which works with $D_n(\cdot)$ at each step t.

Procedure Modified-SGD

Initialize: $w^{(0)} = 0$
Iterate: for $t = 1, 2, \ldots, n$:
  Find α_t to maximize $-\phi_i^*(-\alpha_t) - \frac{\lambda t}{2} \|w^{(t-1)} + (\lambda t)^{-1} \alpha_t x_t\|_2^2$.
  Let $w^{(t)} = \frac{1}{\lambda t} \sum_{i=1}^{t} \alpha_i x_i$
  return α

We have the following result for the convergence of dual objective:

Theorem 3. *Assume that ϕ_i is L-Lipschitz for all i. In addition, assume that (ϕ_i, x_i) are iid samples from the same distribution for all $i = 1, \ldots, n$. At the end of Procedure Modified-SGD, we have*

$$\mathbb{E}[D(\alpha^*) - D(\alpha)] \leq \frac{2L^2 \log(en)}{\lambda n}.$$

Here the expectation is with respect to the random sampling of $\{(\phi_i, x_i) : i = 1, \ldots, n\}$.

Remark 6. *When λ is relatively large, the convergence rate in Theorem 3 for modified-SGD is better than what we can prove for SDCA. This is because Modified-SGD employs a larger step size at each step t for $D_t(\alpha)$ than the corresponding step size in SDCA for $D(\alpha)$. However, the proof requires us to assume that (ϕ_i, x_i) are randomly drawn from a certain distribution, while this extra randomness assumption is not needed for the convergence of SDCA.*

```
┌─────────────────────────────────────────────────────────────────────┐
│            Procedure SDCA with SGD Initialization                   │
│                                                                     │
│   Stage 1: call Procedure Modified-SGD and obtain α                 │
│   Stage 2: call Procedure SDCA with parameter α^(0) = α             │
│                                                                     │
└─────────────────────────────────────────────────────────────────────┘
```

Theorem 4. *Assume that ϕ_i is L-Lipschitz for all i. In addition, assume that (ϕ_i, x_i) are iid samples from the same distribution for all $i = 1, \ldots, n$. Consider Procedure SDCA with SGD Initialization. To obtain a duality gap of $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon_P$ at Stage 2, it suffices to have a total number of SDCA iterations of*

$$T \geq T_0 + n + \frac{4\,L^2}{\lambda\epsilon_P} \geq \lceil n\log(\log(en)) \rceil + n + \frac{20\,L^2}{\lambda\epsilon_P}\,.$$

Moreover, when $t \geq T_0$, we have duality sub-optimality bound of $\mathbb{E}[D(\alpha^*) - D(\alpha^{(t)})] \leq \epsilon_P/2$.

Remark 7. *For Lipschitz loss, ideally we would like to have a computational complexity of $O(n + L^2/(\lambda\epsilon_P))$. Theorem 4 shows that SDCA with SGD at first epoch can achieve no worst than $O(n\log(\log n) + L^2/(\lambda\epsilon_P))$, which is very close to the ideal bound. The result is better than that of vanilla SDCA in Theorem 1 when λ is relatively large, which shows a complexity of $O(n\log(n) + L^2/(\lambda\epsilon_P))$. The difference is caused by small step-sizes in the vanilla SDCA, and its negative effect can be observed in practice. That is, the vanilla SDCA tends to have a slower convergence rate than SGD in the first few iterations when λ is relatively large.*

Remark 8. *Similar to Remark 2, the constant 4 in Theorem 4 can be reduced to 1, and the constant 20 can be reduced to 5.*

5 Refined Analysis for Almost Smooth Loss

Our analysis shows that for smooth loss, SDCA converges faster than SGD (linear versus sub-linear convergence). For non-smooth loss, the analysis does not show any advantage of SDCA over SGD. This does not explain the practical observation that SDCA converges faster than SGD asymptotically even for SVM. This section tries to refine the analysis for Lipschitz loss and shows potential advantage of SDCA over SGD asymptotically.

Although we note that for SVM, Luo and Tseng's analysis [7] shows linear convergence of the form $(1 - \nu)^k$ for dual sub-optimality after k passes over the data, as we mentioned, ν is proportional to the smallest nonzero eigenvalue of the data Gram matrix $X^\top X$, and hence can be arbitrarily bad when two data points $x_i \neq x_j$ becomes very close to each other. Our analysis uses a completely different argument that avoids this dependency on the data Gram matrix.

The analysis relies on the following refined dual strong convexity condition.

Definition 2. *For each i, we define $\gamma_i(\cdot) \geq 0$ so that for all dual variables a and b, and $u \in \partial\phi_i^*(-b)$, we have*

$$\phi_i^*(-a) - \phi_i^*(-b) + u(a - b) \geq \gamma_i(u)|a - b|^2. \tag{4}$$

For the SVM loss, we have $\phi_i(u) = \max(0, 1 - uy_i)$, and $\phi_i^*(-a) = -ay_i$, with $ay_i \in [0, 1]$ and $y_i \in \{\pm 1\}$. It follows that

$$\phi_i^*(-a) - \phi_i^*(-b) + u(a - b) = (b - a)y_i + u(a - b) = |uy_i - 1||a - b| \geq |uy_i - 1| \cdot |a - b|^2.$$

Therefore we may take $\gamma_i(u) = |uy_i - 1|$.

For the absolute deviation loss, we have $\phi_i(u) = |u - y_i|$, and $\phi^*(-a) = -ay_i$ with $a \in [-1, 1]$. It follows that $\gamma_i(u) = |u - y_i|$.

Proposition 1. *Under the assumption of (4). Let* $\gamma_i = \gamma_i(w^{*\top} x_i)$, *we have the following dual strong convexity inequality:*

$$D(\alpha^*) - D(\alpha) \geq \frac{1}{n} \sum_{i=1}^{n} \gamma_i |\alpha_i - \alpha_i^*|^2 + \frac{\lambda}{2}(w - w^*)^\top(w - w^*). \tag{5}$$

Moreover, given $w \in \mathbb{R}^d$ *and* $-a_i \in \partial\phi_i(w^\top x_i)$, *we have*

$$|(w^* - w)^\top x_i| \geq \gamma_i |a_i - \alpha_i^*|. \tag{6}$$

For SVM, we can take $\gamma_i = |w^{*\top} x_i y_i - 1|$, and for the absolute deviation loss, we may take $\gamma_i = |w^{*\top} x_i - y_i|$. Although some of γ_i can be close to zero, in practice, most γ_i will be away from zero, which means $D(\alpha)$ is strongly convex at nearly all points. Under this assumption, we may establish a convergence result for the dual sub-optimality.

Theorem 5. *Consider Procedure SDCA with* $\alpha^{(0)} = 0$. *Assume that* ϕ_i *is* L-*Lipschitz for all* i *and it satisfies (5). Define* $N(u) = \#\{i : \gamma_i < u\}$. *To obtain a dual-suboptimality of* $\mathbb{E}[D(\alpha^*) - D(\alpha^t)] \leq \epsilon_D$, *it suffices to have a total number of iterations of*

$$t \geq 2(n/s)\log(2/\epsilon_D),$$

where $s \in [0, 1]$ *satisfies* $\epsilon_D \geq 8L^2(s/\lambda n)N(s/\lambda n)/n$.

Remark 9. *if* $N(s/\lambda n)/n$ *is small, then Theorem 5 is superior to Theorem 1 for the convergence of the dual objective function. We consider three scenarios. The first scenario is when* $s = 1$. *If* $N(1/\lambda n)/n$ *is small, and* $\epsilon_D \geq 8L^2(1/\lambda n)N(1/\lambda n)/n$, *then the convergence is linear. The second scenario is when there exists* s_0 *so that* $N(s_0/\lambda n) = 0$ *(for SVM, it means that* $\lambda n|w^{*\top} x_i y_i - 1| \geq s_0$ *for all* i*), and since* $\epsilon_D \geq 8L^2(s_0/\lambda n)N(s_0/\lambda n)/n$, *we again have a linear convergence of* $(2n/s_0)\log(2/\epsilon_D)$. *In the third scenario, we assume that* $N(s/\lambda n)/n = O[(s/\lambda n)^\nu]$ *for some* $\nu > 0$, *we can take* $\epsilon_D = O((s/\lambda n)^{1+\nu})$ *and obtain*

$$t \geq O(\lambda^{-1}\epsilon_D^{-1/(1+\nu)}\log(2/\epsilon_D)).$$

The $\log(1/\epsilon_D)$ *factor can be removed in this case with a slightly more complex analysis. This result is again superior to Theorem 1 for dual convergence.*

The following result shows fast convergence of duality gap using Theorem 5.

Theorem 6. *Consider Procedure SDCA with* $\alpha^{(0)} = 0$. *Assume that* ϕ_i *is* L-*Lipschitz for all* i *and it satisfies (4). Let* $\rho \leq 1$ *be the largest eigenvalue of the matrix* $n^{-1}\sum_{i=1} x_i x_i^\top$. *Define* $N(u) = \#\{i : \gamma_i < u\}$. *Assume that at time* $T_0 \geq n$, *we have dual suboptimality of* $\mathbb{E}[D(\alpha^*) - D(\alpha^{(T_0)})] \leq \epsilon_D$, *and define*

$$\tilde{\epsilon}_P = \inf_{\gamma > 0} \left[\frac{N(\gamma)}{n} 4L^2 + \frac{2\epsilon_D}{\min(\gamma, \lambda\gamma^2/(2\rho))} \right],$$

then at time $T = 2T_0$, *we have*

$$\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon_D + \frac{\tilde{\epsilon}_P}{2\lambda T_0}.$$

If for some γ, $N(\gamma)/n$ is small, then Theorem 6 is superior to Theorem 1. In particular, if for some $\gamma > 0$ we have $N(\gamma) = 0$, then

$$\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] = O(\epsilon_D).$$

In this situation, the convergence rate for duality gap is linear as implied by the linear convergence of ϵ_D in Theorem 5.

6 Examples

We will specify the SDCA algorithms for a few common loss functions. For simplicity, we only specify the algorithms without SGD initialization. In practice, instead of complete randomization, we may also run in epochs, and each epoch employs a random permutation of the data. We call this variant SDCA-PERM.

Procedure SDCA−PERM($\alpha^{(0)}$)

Let $w^{(0)} = w(\alpha^{(0)})$
Let $t = 0$
Iterate: for epoch $k = 1, 2, \ldots$
  Let $\{i_1, \ldots, i_n\}$ be a random permutation of $\{1, \ldots, n\}$
  Iterate: for $j = 1, 2, \ldots, n$:
    $t \leftarrow t + 1$
    $i = i_j$
    Find $\Delta\alpha_i$ to increase dual   (*)
    $\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$
    $w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1} \Delta\alpha_i x_i$
Output (Averaging option):
  Let $\bar{\alpha} = \frac{1}{T-T_0} \sum_{i=T_0+1}^{T} \alpha^{(t-1)}$
  Let $\bar{w} = w(\bar{\alpha}) = \frac{1}{T-T_0} \sum_{i=T_0+1}^{T} w^{(t-1)}$
  return $\bar{w}$
Output (Random option):
  Let $\bar{\alpha} = \alpha^{(t)}$ and $\bar{w} = w^{(t)}$ for some random $t \in T_0 + 1, \ldots, T$
  return $\bar{w}$

Lipschitz loss

Hinge loss is used in SVM. We have $\phi_i(u) = \max\{0, 1 - y_i u\}$ and $\phi_i^*(-a) = -a y_i$ with $a y_i \in [0, 1]$. Absolute deviation loss is used in quantile regression. We have $\phi_i(u) = |u - y_i|$ and $\phi_i^*(-a) = -a y_i$ with $a \in [-1, 1]$.

For the hinge loss, step (*) in Procedure SDCA-PERM has a closed form solution as

$$\Delta\alpha_i = y_i \max\left(0, \min\left(1, \frac{1 - x_i^\top w^{(t-1)} y_i}{\|x_i\|_2^2/(\lambda n)} + \alpha_i^{(t-1)} y_i\right)\right) - \alpha_i^{(t-1)}.$$

For absolute deviation loss, step (*) in Procedure SDCA-PERM has a closed form solution as

$$\Delta\alpha_i = \max\left(-1, \min\left(1, \frac{y_i - x_i^\top w^{(t-1)}}{\|x_i\|_2^2/(\lambda n)} + \alpha_i^{(t-1)}\right)\right) - \alpha_i^{(t-1)}.$$

Both hinge loss and absolute deviation loss are 1-Lipschitz. Therefore, we expect a convergence behavior of no worse than

$$O\left(n\ln n + \frac{1}{\lambda\epsilon}\right)$$

without SGD initialization based on Theorem 1. The refined analysis in Section 5 suggests a rate that can be significantly better, and this is confirmed with our empirical experiments.

Smooth loss

Squared loss is used in ridge regression. We have $\phi_i(u) = (u - y_i)^2$, and $\phi_i^*(-a) = -ay_i + a^2/4$. Log loss is used in logistic regression. We have $\phi_i(u) = \log(1 + \exp(-y_i u))$, and $\phi_i^*(-a) = ay_i \ln(ay_i) + (1 - ay_i)\ln(1 - ay_i)$ with $ay_i \in [0, 1]$.

For squared loss, step (*) in Procedure SDCA-PERM has a closed form solution as

$$\Delta\alpha_i = \frac{y_i - x_i^\top w^{(t-1)} - 0.5\alpha_i^{(t-1)}}{0.5 + \|x_i\|_2^2/(\lambda n)}.$$

For log loss, step (*) in Procedure SDCA-PERM does not have a closed form solution. However, one may start with the approximate solution,

$$\Delta\alpha_i = \frac{(1 + \exp(x_i^\top w^{(t-1)} y_i))^{-1} y_i - \alpha_i^{(t-1)}}{\max(1, 0.25 + \|x_i\|_2^2/(\lambda n))},$$

and further use several steps of Newton's update to get a more accurate solution.

Finally, we present a smooth variant of the hinge-loss, as defined below. Recall that the hinge loss function (for positive labels) is $\phi(u) = \max\{0, 1 - u\}$ and we have $\phi^*(-a) = -a$ with $a \in [0, 1]$. Consider adding to ϕ^* the term $\frac{\gamma}{2}a^2$ which yields the γ-strongly convex function

$$\tilde{\phi}_\gamma^*(a) = \phi^*(a) + \frac{\gamma}{2}a^2 .$$

Then, its conjugate, which is defined below, is $(1/\gamma)$-smooth. We refer to it as the *smoothed hinge-loss* (for positive labels):

$$\tilde{\phi}_\gamma(x) = \max_{a\in[-1,0]} \left[ax - a - \frac{\gamma}{2}a^2\right]$$
$$= \begin{cases} 0 & x > 1 \\ 1 - x - \gamma/2 & x < 1 - \gamma \\ \frac{1}{2\gamma}(1 - x)^2 & \text{o.w.} \end{cases} \tag{7}$$

For the smoothed hinge loss, step (*) in Procedure SDCA-PERM has a closed form solution as

$$\Delta\alpha_i = y_i \max\left(0, \min\left(1, \frac{1 - x_i^\top w^{(t-1)} y_i - \gamma\alpha_i^{(t-1)} y_i}{\|x_i\|_2^2/(\lambda n) + \gamma} + \alpha_i^{(t-1)} y_i\right)\right) - \alpha_i^{(t-1)}.$$

Both log loss and squared loss are 1-smooth. The smoothed-hinge loss is $1/\gamma$ smooth. Therefore we expect a convergence behavior of no worse than

$$O\left(\left(n + \frac{1}{\gamma\lambda}\right)\ln\frac{1}{\epsilon}\right).$$

This is confirmed in our empirical experiments.

10

7 Experimental Results

In this section we demonstrate the tightness of our theory. All our experiments are performed with the smooth variant of the hinge-loss defined in (7), where the value of γ is taken from the set $\{0, 0.01, 0.1, 1\}$. Note that for $\gamma = 0$ we obtain the vanilla non-smooth hinge-loss.

In the experiments, we use ϵ_D to denote the dual sub-optimality, and ϵ_P to denote the primal sub-optimality (note that this is different than the notation in our analysis which uses ϵ_P to denote the duality gap). It follows that $\epsilon_D + \epsilon_P$ is the duality gap.

7.1 Data

The experiments were performed on three large datasets with very different feature counts and sparsity, which were kindly provided by Thorsten Joachims. The astro-ph dataset classifies abstracts of papers from the physics ArXiv according to whether they belong in the astro-physics section; CCAT is a classification task taken from the Reuters RCV1 collection; and cov1 is class 1 of the covertype dataset of Blackard, Jock & Dean. The following table provides details of the dataset characteristics.

Dataset	Training Size	Testing Size	Features	Sparsity
astro-ph	29882	32487	99757	0.08%
CCAT	781265	23149	47236	0.16%
cov1	522911	58101	54	22.22%

7.2 Linear convergence for Smooth Hinge-loss

Our first experiments are with ϕ_γ where we set $\gamma = 1$. The goal of the experiment is to show that the convergence is indeed linear. We ran the SDCA algorithm for solving the regularized loss minimization problem with different values of regularization parameter λ. Figure 1 shows the results. Note that a logarithmic scale is used for the vertical axis. Therefore, a straight line corresponds to linear convergence. We indeed observe linear convergence for the duality gap.

7.3 Convergence for non-smooth Hinge-loss

Next we experiment with the original hinge loss, which is 1-Lipschitz but is not smooth. We again ran the SDCA algorithm for solving the regularized loss minimization problem with different values of regularization parameter λ. Figure 2 shows the results. As expected, the overall convergence rate is slower than the case of a smoothed hinge-loss. However, it is also apparent that for large values of λ a linear convergence is still exhibited, as expected according to our refined analysis. The bounds plotted are based on Theorem 1, which are slower than what we observe, as expected from the refined analysis in Section 5.

7.4 Effect of smoothness parameter

We next show the effect of the smoothness parameter. Figure 3 shows the effect of the smoothness parameter on the rate of convergence. As can be seen, the convergence becomes faster as the loss function becomes smoother. However, the difference is more dominant when λ decreases.

Figure 4 shows the effect of the smoothness parameter on the zero-one test error. It is noticeable that the difference in zero-one test error for different values of γ is not very significant, and differs from data to
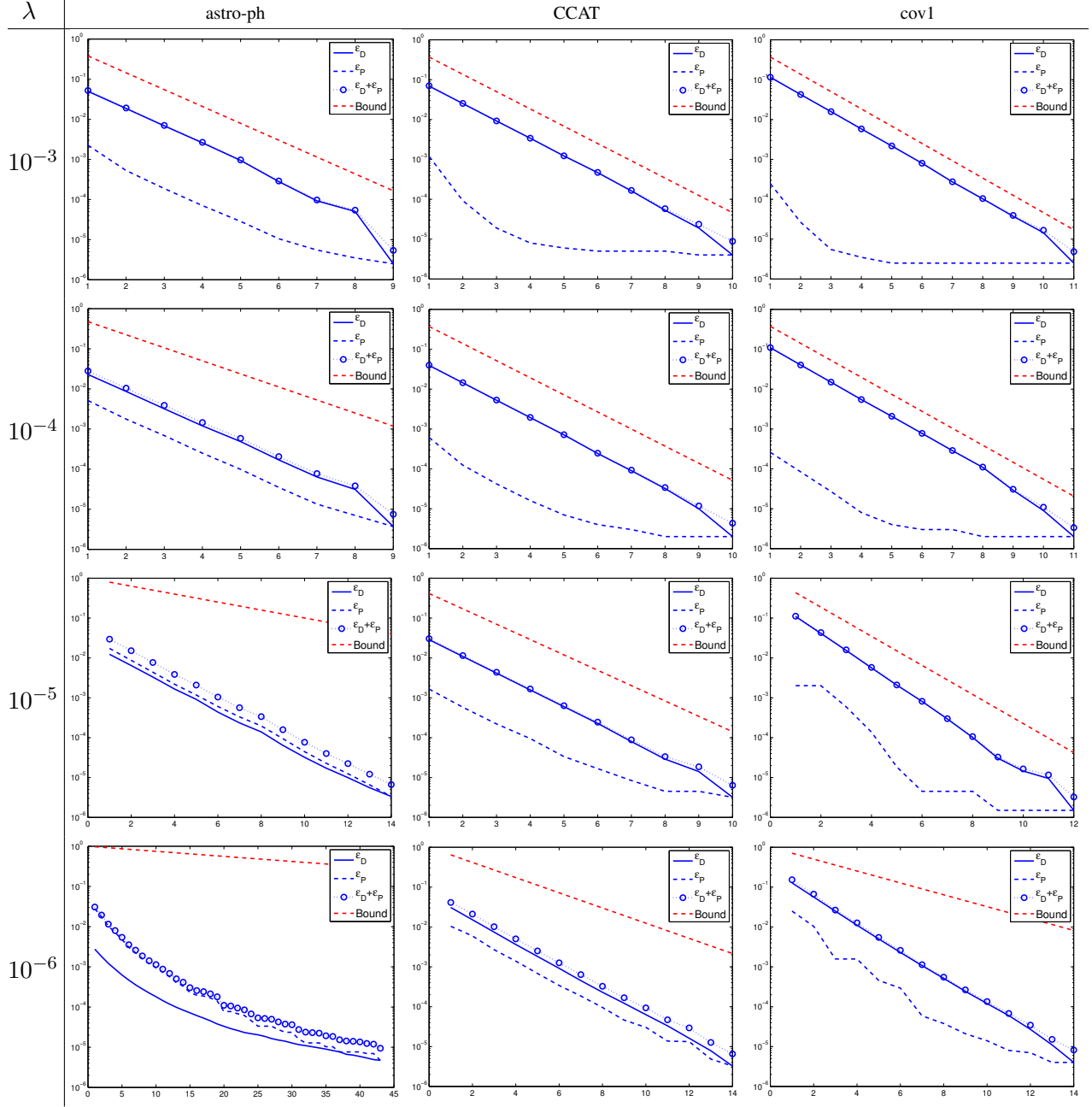
Figure 1: Experiments with the smoothed hinge-loss ($\gamma = 1$). The primal and dual sub-optimality, the duality gap, and our bound are depicted as a function of the number of epochs, on the astro-ph (left), CCAT (center) and cov1 (right) datasets. In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).
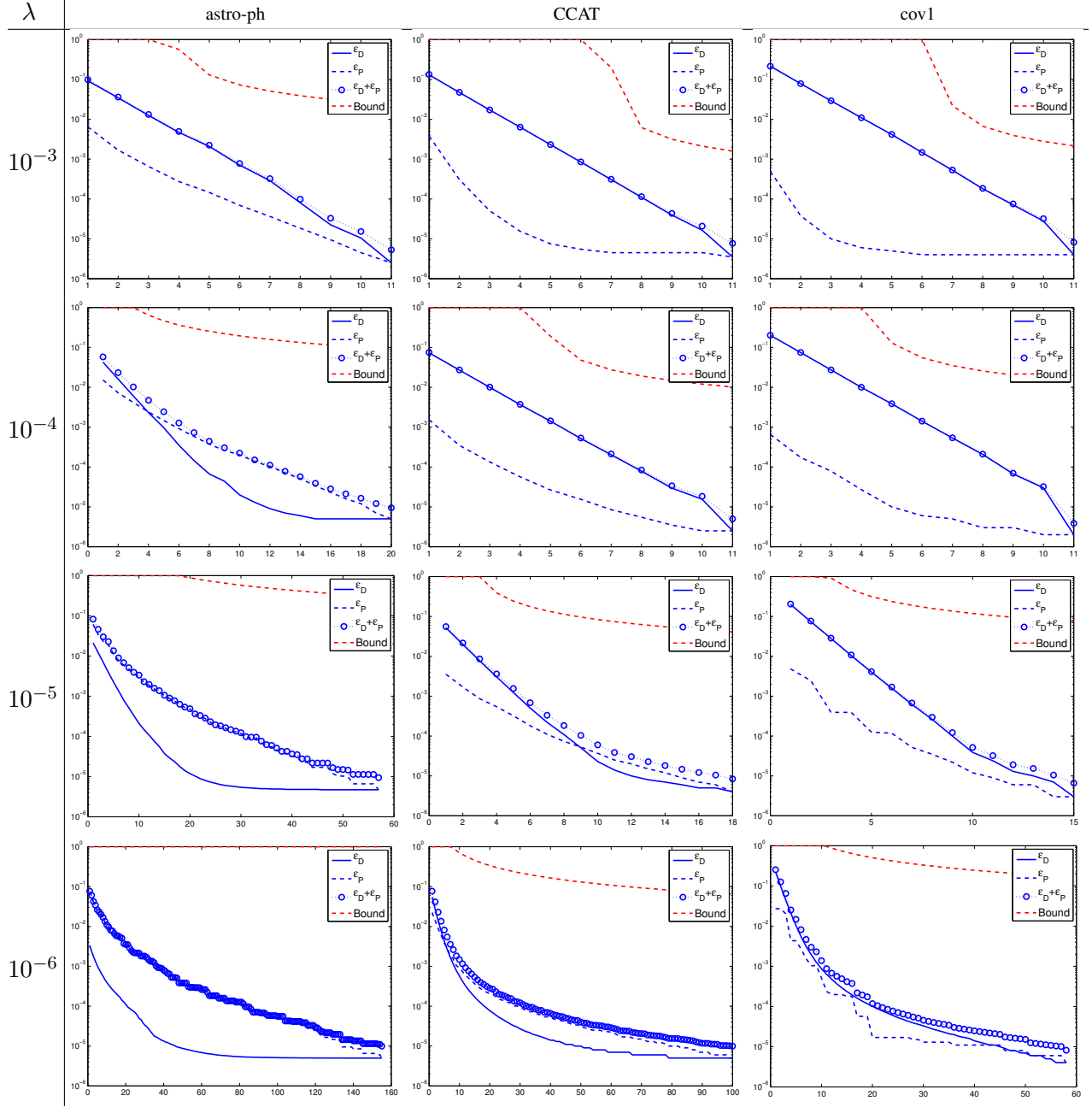
Figure 2: Experiments with the hinge-loss (non-smooth). The primal and dual sub-optimality, the duality gap, and our bound are depicted as a function of the number of epochs, on the astro-ph (left), CCAT (center) and cov1 (right) datasets. In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).

data. Since in terms of runtime a larger value of γ is preferable, this might suggest that the smooth versions of the hinge-loss are preferable.

7.5 Cyclic vs. Stochastic vs. Random Permutation

In Figure 5 we compare choosing dual variables at random with repetitions (as done in SDCA) vs. choosing dual variables using a random permutation at each epoch (as done in SDCA-Perm) vs. choosing dual variables in a fixed cyclic order (that was chosen once at random). As can be seen, a cyclic order does not lead to linear convergence and yields actual convergence rate much slower than the other methods and even worse than our bound. As mentioned before, some of the earlier analyses such as [7] can be applied both to stochastic and to cyclic dual coordinate ascent methods with similar results. This means that their analysis, which can be no better than the behavior of cyclic dual coordinate ascent, is inferior to our analysis. Finally, we also observe that SDCA-Perm is sometimes faster than SDCA.

7.6 Comparison to SGD

We next compare SDCA to Stochastic Gradient Descent (SGD). One clear advantage of SDCA is the availability of a clear stopping condition (by calculating the duality gap). In Figure 6 and Figure 7 we present the primal sub-optimality of SDCA, SDCA-Perm, and SGD. As can be seen, SDCA converges faster than SGD in most regimes. SGD can be better if both λ is high and one performs a very small number of epochs. This is in line with our theory of Section 4. However, SDCA quickly catches up.

8 Proofs

For convenience, we list the following simple facts about primal and dual formulations, which will used in the proofs. For each i, we have

$$-\alpha_i^* \in \partial \phi_i(w^{*\top} x_i), \quad w^{*\top} x_i \in \partial \phi_i^*(-\alpha_i^*),$$

and

$$w^* = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i^* x_i.$$

8.1 Proof of Theorem 2

The key lemma is the following:

Lemma 1. *Assume that ϕ_i^* is γ-strongly-convex (where γ can be zero). Then, for any iteration t and any $s \in [0, 1]$ we have*

$$\mathbb{E}[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] \geq \frac{s}{n} \, \mathbb{E}[P(w^{(t-1)}) - D(\alpha^{(t-1)})] - \left(\frac{s}{n}\right)^2 \frac{G^{(t)}}{2\lambda} \, ,$$

where

$$G^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \left(\|x_i\|^2 - \frac{\gamma(1-s)\lambda n}{s} \right) \mathbb{E}[(u_i^{(t-1)} - \alpha_i^{(t-1)})^2],$$

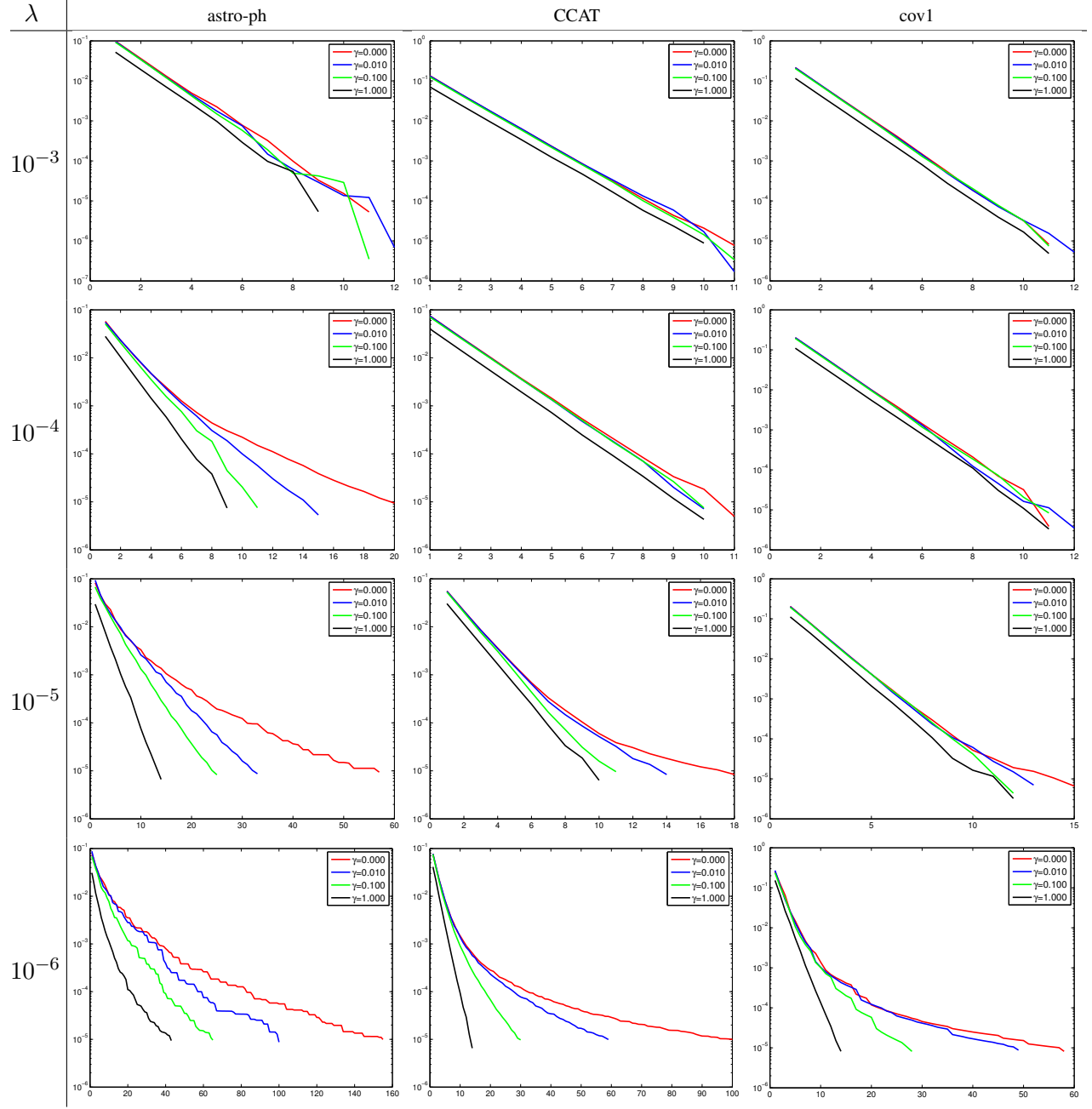*and $-u_i^{(t-1)} \in \partial \phi_i(x_i^\top w^{(t-1)})$.*

14

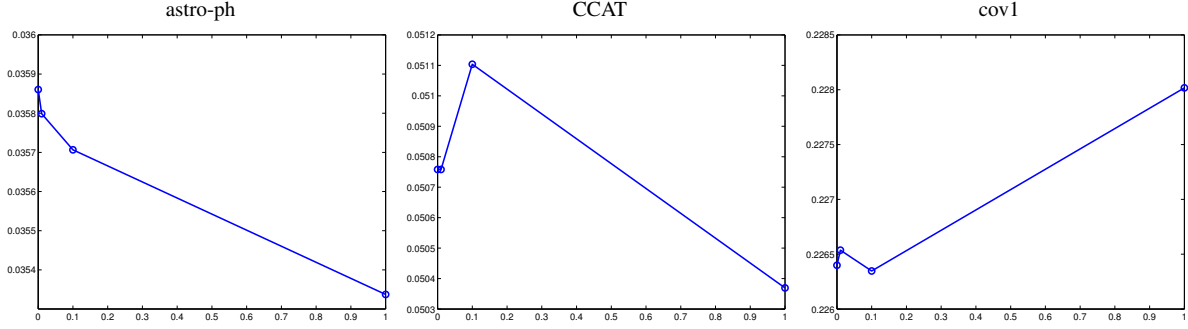Figure 3: Duality gap as a function of the number of rounds for different values of γ.

Figure 4: Zero-one test error as a function of the value of γ.

Proof. Since only the i'th element of α is updated, the improvement in the dual objective can be written as

$$n[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] = \underbrace{\left(-\phi^*(-\alpha_i^{(t)}) - \frac{\lambda n}{2}\|w^{(t)}\|^2\right)}_{A} - \underbrace{\left(-\phi^*(-\alpha_i^{(t-1)}) - \frac{\lambda n}{2}\|w^{(t-1)}\|^2\right)}_{B}$$

By the definition of the update we have for all $s \in [0,1]$ that

$$A = \max_{\Delta\alpha_i} -\phi^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i\|^2$$

$$\geq -\phi^*(-(\alpha_i^{(t-1)} + s(u_i^{(t)} - \alpha_i^{(t-1)}))) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}s(u_i^{(t)} - \alpha_i^{(t-1)})x_i\|^2 \quad (8)$$

From now on, we omit the superscripts and subscripts. Since ϕ^* is γ-strongly convex, we have that

$$\phi^*(-(\alpha + s(u - \alpha))) = \phi^*(s(-u) + (1-s)(-\alpha)) \leq s\phi^*(-u) + (1-s)\phi^*(-\alpha) - \frac{\gamma}{2}s(1-s)(u-\alpha)^2 \quad (9)$$

Combining this with (8) and rearranging terms we obtain that

$$A \geq -s\phi^*(-u) - (1-s)\phi^*(-\alpha) + \frac{\gamma}{2}s(1-s)(u-\alpha)^2 - \frac{\lambda n}{2}\|w + (\lambda n)^{-1}s(u-\alpha)x\|^2$$

$$= -s\phi^*(-u) - (1-s)\phi^*(-\alpha) + \frac{\gamma}{2}s(1-s)(u-\alpha)^2 - \frac{\lambda n}{2}\|w\|^2 - s(u-\alpha)w^\top x - \frac{s^2(u-\alpha)^2}{2\lambda n}\|x\|^2$$

$$= \underbrace{-s(\phi^*(-u) + uw^\top x)}_{s\,\phi(w^T x)} + \underbrace{(-\phi^*(-\alpha) - \frac{\lambda n}{2}\|w\|^2)}_{B} + \frac{s}{2}\left(\gamma(1-s) - \frac{s\|x\|^2}{\lambda n}\right)(u-\alpha)^2 + s(\phi^*(-\alpha) + \alpha w^\top x),$$

where we used $-u \in \partial\phi(w^\top x)$ which yields $\phi^*(-u) = -uw^\top x - \phi(w^\top x)$. Therefore

$$A - B \geq s\left[\phi(w^T x) + \phi^*(-\alpha) + \alpha w^\top x + \left(\frac{\gamma(1-s)}{2} - \frac{s\|x\|^2}{2\lambda n}\right)(u-\alpha)^2\right]. \quad (10)$$

Next note that

$$P(w) - D(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(w^\top x_i) + \frac{\lambda}{2}w^\top w - \left(-\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(-\alpha_i) - \frac{\lambda}{2}w^\top w\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\phi_i(w^\top x_i) + \phi_i^*(-\alpha_i) + \alpha_i w^\top x_i\right)$$
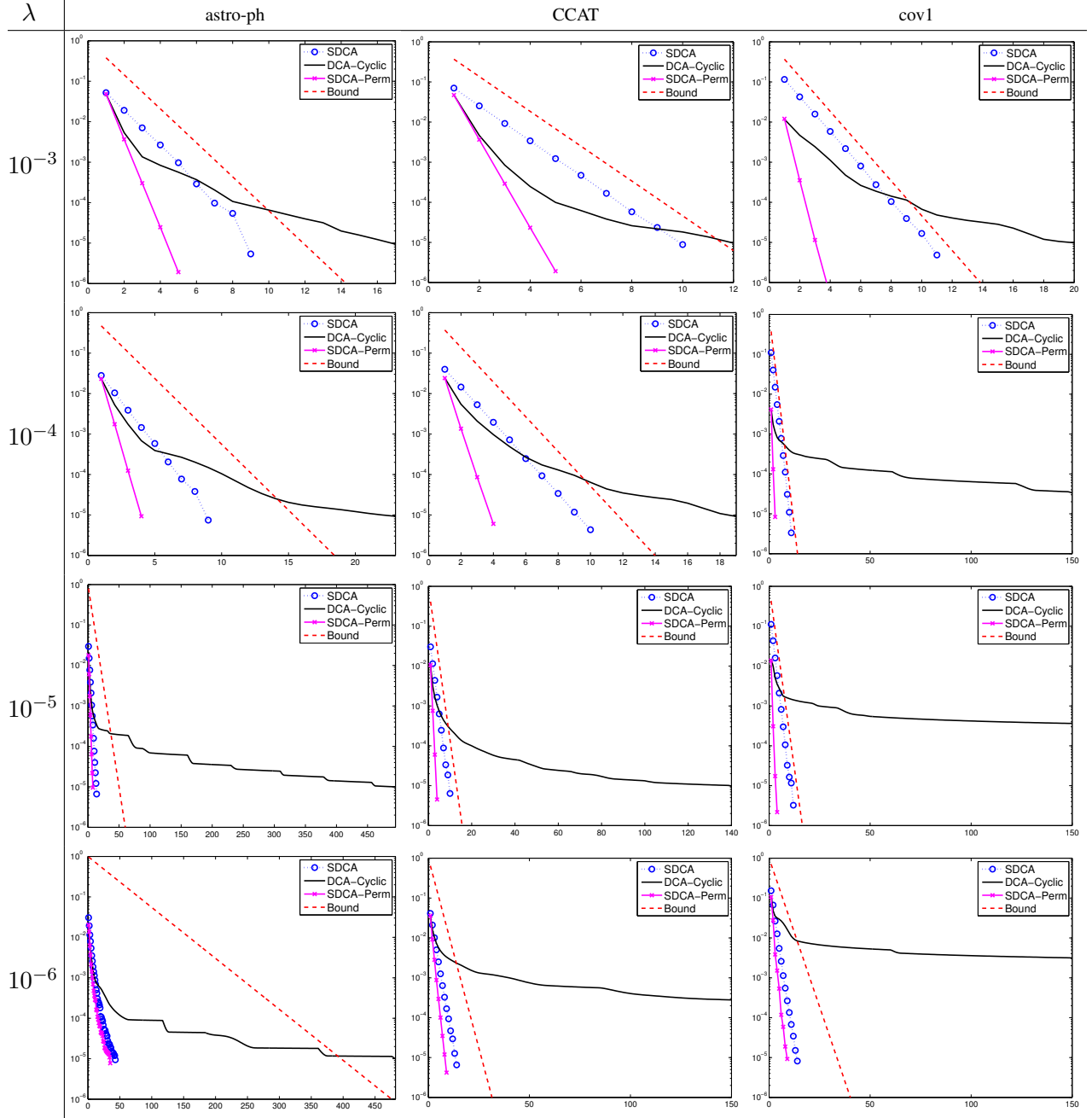
16

Figure 5: Comparing the duality gap achieved by choosing dual variables at random with repetitions (SDCA), choosing dual variables at random without repetitions (SDCA-Perf), or using a fixed cyclic order. In all cases, the duality gap is depicted as a function of the number of epochs for different values of λ. The loss function is the smooth hinge loss with $\gamma = 1$.
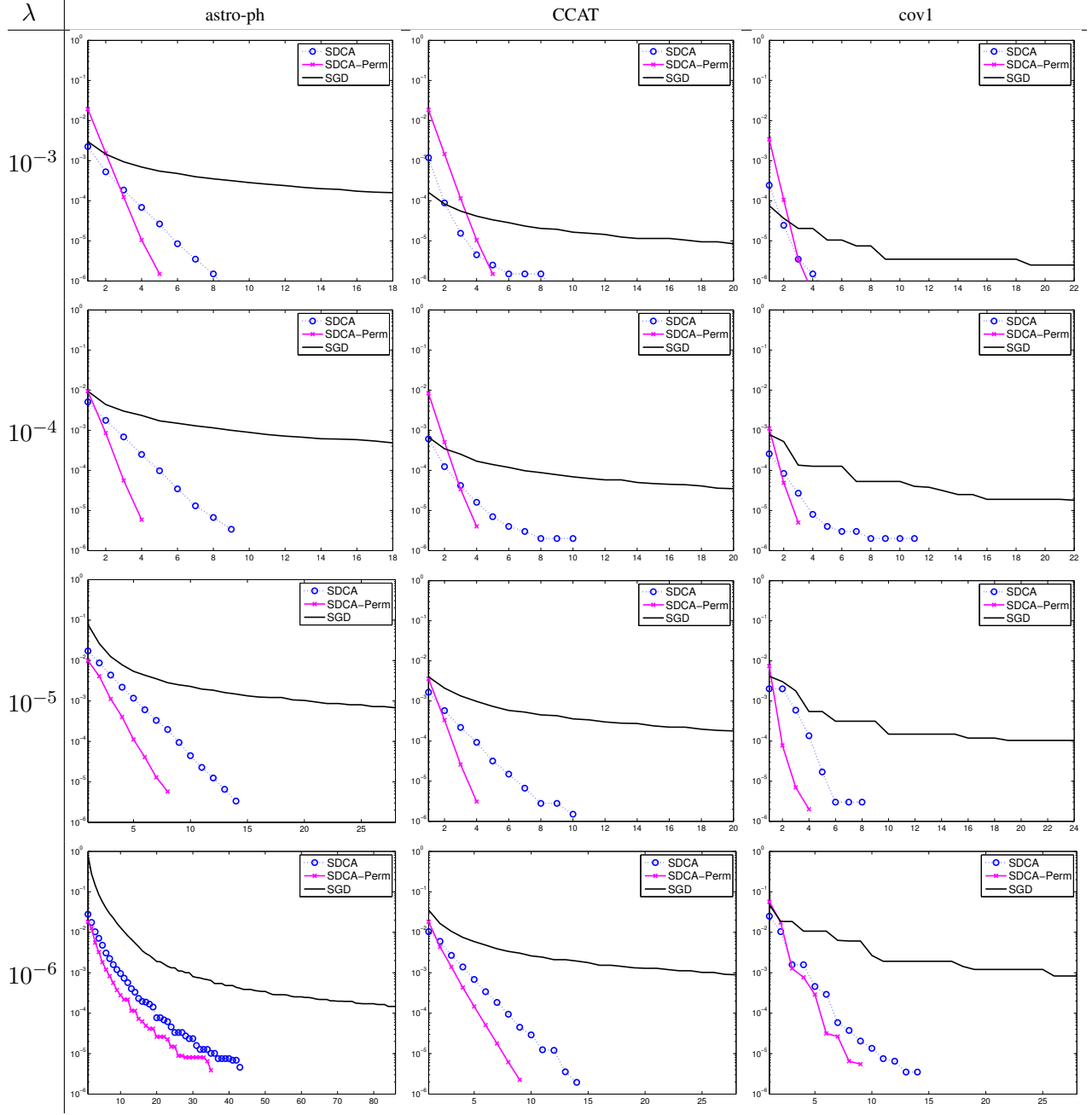
Figure 6: Comparing the primal sub-optimality of SDCA and SGD for the smoothed hinge-loss ($\gamma = 1$). In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).

Figure 7: Comparing the primal sub-optimality of SDCA and SGD for the non-smooth hinge-loss ($\gamma = 0$). In all plots the horizontal axis is the number of iterations divided by training set size (corresponding to the number of epochs through the data).
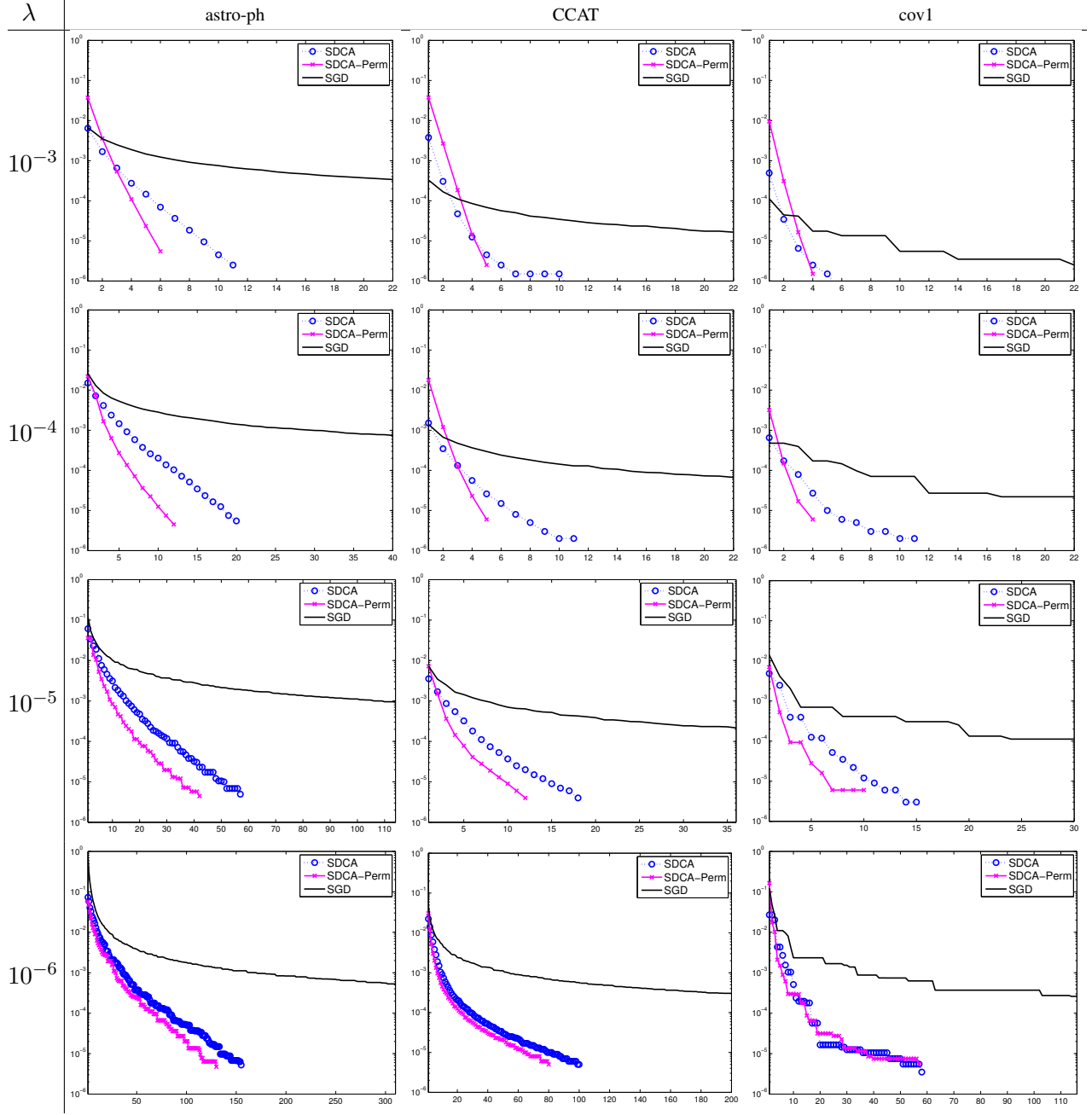
19

Therefore, if we take expectation of (10) w.r.t. the choice of i we obtain that

$$\frac{1}{s} \, \mathbb{E}[A - B] \geq \mathbb{E}[P(w) - D(\alpha)] - \frac{s}{2\lambda n} \cdot \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(\|x_i\|^2 - \frac{\gamma(1-s)\lambda n}{s} \right) (u_i - \alpha_i)^2}_{=G^{(t)}} .$$

We have obtained that

$$\frac{n}{s} \, \mathbb{E}[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] \geq \mathbb{E}[P(w^{(t-1)}) - D(\alpha^{(t-1)})] - \frac{s \, G^{(t)}}{2\lambda n} . \qquad (11)$$

Multiplying both sides by s/n concludes the proof of the lemma. $\qquad\square$

We also use the following simple lemma:

Lemma 2. *For all α, $D(\alpha) \leq P(w^*) \leq P(0) \leq 1$. In addition, $D(0) \geq 0$.*

Proof. The first inequality is by weak duality, the second is by the optimality of w^*, and the third by the assumption that $\phi_i(0) \leq 1$. For the last inequality we use $-\phi_i^*(0) = -\max_z(0 - \phi_i(z)) = \min_z \phi_i(z) \geq 0$, which yields $D(0) \geq 0$. $\qquad\square$

Equipped with the above lemmas we are ready to prove Theorem 2.

Proof of Theorem 2. The assumption that ϕ_i is $(1/\gamma)$-smooth implies that ϕ_i^* is γ-strongly-convex. We will apply Lemma 1 with $s = \frac{\lambda n \gamma}{1 + \lambda n \gamma} \in [0, 1]$. Recall that $\|x_i\| \leq 1$. Therefore, the choice of s implies that $\|x_i\|^2 - \frac{\gamma(1-s)\lambda n}{s} \leq 0$, and hence $G^{(t)} \leq 0$ for all t. This yields,

$$\mathbb{E}[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] \geq \frac{s}{n} \, \mathbb{E}[P(w^{(t-1)}) - D(\alpha^{(t-1)})] .$$

But since $\epsilon_D^{(t-1)} := D(\alpha^*) - D(\alpha^{(t-1)}) \leq P(w^{(t-1)}) - D(\alpha^{(t-1)})$ and $D(\alpha^{(t)}) - D(\alpha^{(t-1)}) = \epsilon_D^{(t-1)} - \epsilon_D^{(t)}$, we obtain that

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \left(1 - \tfrac{s}{n}\right) \mathbb{E}[\epsilon_D^{(t-1)}] \leq \left(1 - \tfrac{s}{n}\right)^t \mathbb{E}[\epsilon_D^{(0)}] \leq \left(1 - \tfrac{s}{n}\right)^t \leq \exp(-st/n) = \exp\left(-\frac{\lambda \gamma t}{1 + \lambda \gamma n}\right) .$$

This would be smaller than ϵ_D if

$$t \geq \left(n + \tfrac{1}{\lambda \gamma}\right) \log(1/\epsilon_D) .$$

It implies that

$$\mathbb{E}[P(w^{(t)}) - D(\alpha^{(t)})] \leq \frac{n}{s} \, \mathbb{E}[\epsilon_D^{(t)} - \epsilon_D^{(t+1)}] \leq \frac{n}{s} \, \mathbb{E}[\epsilon_D^{(t)}]. \qquad (12)$$

So, requiring $\epsilon_D^{(t)} \leq \frac{s}{n} \epsilon_P$ we obtain a duality gap of at most ϵ_P. This means that we should require

$$t \geq \left(n + \tfrac{1}{\lambda \gamma}\right) \log\left((n + \tfrac{1}{\lambda \gamma}) \cdot \tfrac{1}{\epsilon_P}\right) ,$$

which proves the first part of Theorem 2.

Next, we sum (12) over $t = T_0, \dots, T - 1$ to obtain

$$\mathbb{E}\left[\frac{1}{T - T_0} \sum_{t=T_0}^{T-1} (P(w^{(t)}) - D(\alpha^{(t)}))\right] \leq \frac{n}{s(T - T_0)} \mathbb{E}[D(\alpha^{(T)}) - D(\alpha^{(T_0)})].$$

Now, if we choose $\bar{w}, \bar{\alpha}$ to be either the average vectors or a randomly chosen vector over $t \in \{T_0 + 1, \dots, T\}$, then the above implies

$$\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \frac{n}{s(T - T_0)} \mathbb{E}[D(\alpha^{(T)}) - D(\alpha^{(T_0)})] \leq \frac{n}{s(T - T_0)} \mathbb{E}[\epsilon_D^{(T_0)}].$$

It follows that in order to obtain a result of $\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \epsilon_P$, we only need to have

$$\mathbb{E}[\epsilon_D^{(T_0)}] \leq \frac{s(T - T_0)\epsilon_P}{n} = \frac{(T - T_0)\epsilon_P}{n + \frac{1}{\lambda\gamma}}.$$

This implies the second part of Theorem 2, and concludes the proof. $\qquad\square$

8.2 Proof of Theorem 1

Next, we turn to the case of Lipschitz loss function. We rely on the following lemma.

Lemma 3. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz function. Then, for any α s.t. $|\alpha| > L$ we have that $\phi^*(\alpha) = \infty$.*

Proof. Fix some $\alpha > L$. By definition of the conjugate we have

$$\begin{aligned}
\phi^*(\alpha) &= \sup_x [\alpha\, x - \phi(x)] \\
&\geq -\phi(0) + \sup_x [\alpha\, x - (\phi(x) - \phi(0))] \\
&\geq -\phi(0) + \sup_x [\alpha\, x - L|x - 0|] \\
&\geq -\phi(0) + \sup_{x > 0}(\alpha - L)\, x = \infty \;.
\end{aligned}$$

Similar argument holds for $\alpha < -L$. $\qquad\square$

A direct corollary of the above lemma is:

Lemma 4. *Suppose that for all i, ϕ_i is L-Lipschitz. Let $G^{(t)}$ be as defined in Lemma 1 (with $\gamma = 0$). Then, $G^{(t)} \leq 4\,L^2$.*

Proof. Using Lemma 3 we know that $|\alpha_i^{(t-1)}| \leq L$, and in addition by the relation of Lipschitz and subgradients we have $|u_i^{(t-1)}| \leq L$. Thus, $(u_i^{(t-1)} - \alpha_i^{(t-1)})^2 \leq 4L^2$, and the proof follows. $\qquad\square$

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let $G = \max_t G^{(t)}$ and note that by Lemma 4 we have $G \leq 4L^2$. Lemma 1, with $\gamma = 0$, tells us that

$$\mathbb{E}[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] \geq \frac{s}{n}\, \mathbb{E}[P(w^{(t-1)}) - D(\alpha^{(t-1)})] - \left(\frac{s}{n}\right)^2 \frac{G}{2\lambda}, \qquad (13)$$

which implies that

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \left(1 - \tfrac{s}{n}\right) \mathbb{E}[\epsilon_D^{(t-1)}] + \left(\tfrac{s}{n}\right)^2 \tfrac{G}{2\lambda} \;.$$

21

We next show that the above yields

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \frac{2G}{\lambda(2n + t - t_0)} \tag{14}$$

for all $t \geq t_0 = \max(0, \lceil n \log(2\lambda n \epsilon_D^{(0)}/G) \rceil)$. Indeed, let us choose $s = 1$, then at $t = t_0$, we have

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \left(1 - \tfrac{1}{n}\right)^t \epsilon_D^{(0)} + \frac{G}{2\lambda n^2} \frac{1}{1 - (1 - 1/n)} \leq e^{-t/n} \epsilon_D^{(0)} + \frac{G}{2\lambda n} \leq \frac{G}{\lambda n} \,.$$

This implies that (14) holds at $t = t_0$. For $t > t_0$ we use an inductive argument. Suppose the claim holds for $t - 1$, therefore

$$\mathbb{E}[\epsilon_D^{(t)}] \leq \left(1 - \tfrac{s}{n}\right) \mathbb{E}[\epsilon_D^{(t-1)}] + \left(\tfrac{s}{n}\right)^2 \frac{G}{2\lambda} \leq \left(1 - \tfrac{s}{n}\right) \frac{2G}{\lambda(2n + t - 1 - t_0)} + \left(\tfrac{s}{n}\right)^2 \frac{G}{2\lambda}.$$

Choosing $s = 2n/(2n - t_0 + t - 1) \in [0, 1]$ yields

$$\begin{aligned}
\mathbb{E}[\epsilon_D^{(t)}] &\leq \left(1 - \tfrac{2}{2n - t_0 + t - 1}\right) \frac{2G}{\lambda(2n - t_0 + t - 1)} + \left(\tfrac{2}{2n - t_0 + t - 1}\right)^2 \frac{G}{2\lambda} \\
&= \frac{2G}{\lambda(2n - t_0 + t - 1)} \left(1 - \tfrac{1}{2n - t_0 + t - 1}\right) \\
&= \frac{2G}{\lambda(2n - t_0 + t - 1)} \frac{2n - t_0 + t - 2}{2n - t_0 + t - 1} \\
&\leq \frac{2G}{\lambda(2n - t_0 + t - 1)} \frac{2n - t_0 + t - 1}{2n - t_0 + t} \\
&= \frac{2G}{\lambda(2n - t_0 + t)} \,.
\end{aligned}$$

This provides a bound on the dual sub-optimality. We next turn to bound the duality gap. Summing (13) over $t = T_0 + 1, \ldots, T$ and rearranging terms we obtain that

$$\mathbb{E}\left[\frac{1}{T - T_0} \sum_{t = T_0 + 1}^{T} (P(w^{(t-1)}) - D(\alpha^{(t-1)}))\right] \leq \frac{n}{s(T - T_0)} \mathbb{E}[D(\alpha^{(T)}) - D(\alpha^{(T_0)})] + \frac{sG}{2\lambda n}$$

Now, if we choose $\bar{w}, \bar{\alpha}$ to be either the average vectors or a randomly chosen vector over $t \in \{T_0 + 1, \ldots, T\}$, then the above implies

$$\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] \leq \frac{n}{s(T - T_0)} \mathbb{E}[D(\alpha^{(T)}) - D(\alpha^{(T_0)})] + \frac{sG}{2\lambda n} \,.$$

If $T \geq n + T_0$ and $T_0 \geq t_0$, we can set $s = n/(T - T_0)$ and combining with (14) we obtain

$$\begin{aligned}
\mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] &\leq \mathbb{E}[D(\alpha^{(T)}) - D(\alpha^{(T_0)})] + \frac{G}{2\lambda(T - T_0)} \\
&\leq \mathbb{E}[D(\alpha^*) - D(\alpha^{(T_0)})] + \frac{G}{2\lambda(T - T_0)} \\
&\leq \frac{2G}{\lambda(2n - t_0 + T_0)} + \frac{G}{2\lambda(T - T_0)} \,.
\end{aligned}$$

A sufficient condition for the above to be smaller than ϵ_P is that $T_0 \geq \frac{4G}{\lambda \epsilon_P} - 2n + t_0$ and $T \geq T_0 + \frac{G}{\lambda \epsilon_P}$. It also implies that $\mathbb{E}[D(\alpha^*) - D(\alpha^{(T_0)})] \leq \epsilon_P/2$. Since we also need $T_0 \geq t_0$ and $T - T_0 \geq n$, the overall number of required iterations can be

$$T_0 \geq \max\{t_0, 4G/(\lambda \epsilon_P) - 2n + t_0\}, \quad T - T_0 \geq \max\{n, G/(\lambda \epsilon_P)\}.$$

We conclude the proof by noticing that $\epsilon_D^{(0)} \leq 1$ (Lemma 2), which implies that $t_0 \leq \max(0, \lceil n \log(2\lambda n/G) \rceil)$.

$\square$

8.3   Proof of Theorem 3

We assume that (ϕ_t, x_t) are randomly drawn from a distribution D, and define the population optimizer

$$w_D^* = \operatorname*{argmin}_w P_D(w), \qquad P_D(w) = \mathbb{E}_{(\phi,x)\sim D}\left[\phi(w^\top x) + \frac{\lambda}{2}\|w\|^2\right].$$

By definition, we have $P(w^*) \leq P(w_D^*)$ for any specific realization of $\{(\phi_t, x_t) : t = 1, \ldots, n\}$. Therefore

$$\mathbb{E}\, P(w^*) \leq \mathbb{E}\, P(w_D^*) = \mathbb{E}\, P_D(w_D^*),$$

where the expectation is with respect to the choice of examples, and note that both $P(\cdot)$ and w^* are sample dependent.

After each step t, we let $\alpha^{(t)} = [\alpha_1, \ldots, \alpha_t]$, and let $-u \in \partial\phi_{t+1}(x_{t+1}^\top w^{(t)})$. We have, for all t,

$$(t+1)D_{t+1}(\alpha^{(t+1)}) - tD_t(\alpha^{(t)}) = -\phi_{t+1}^*(-\alpha_{t+1}^{(t+1)}) - (t+1)\frac{\lambda}{2}\|w^{(t+1)}\|^2 + t\frac{\lambda}{2}\|w^{(t)}\|^2$$

$$= -\phi_{t+1}^*(-\alpha_{t+1}^{(t+1)}) - \frac{1}{2(t+1)\lambda}\|\lambda t w^{(t)} + \alpha_{t+1}^{(t+1)} x_{t+1}\|^2 + \frac{1}{2t\lambda}\|\lambda t w^{(t)}\|^2$$

$$\geq -\phi^*(u) - \frac{1}{2(t+1)\lambda}\|\lambda t w^{(t)} + u x_{t+1}\|^2 + \frac{1}{2t\lambda}\|\lambda t w^{(t)}\|^2$$

$$= -\phi_{t+1}^*(u) - \frac{t}{t+1}x_{t+1}^\top w^{(t)} u + \frac{1}{2\lambda}\left(\frac{1}{t} - \frac{1}{t+1}\right)\|\lambda t w^{(t)}\|^2 - \frac{u^2\|x_{t+1}\|^2}{2(t+1)\lambda}$$

$$= -\phi_{t+1}^*(u) - x_{t+1}^\top w^{(t)} u + \left(1 - \frac{t}{t+1}\right)x_{t+1}^\top w^{(t)} u + \frac{1}{2(t+1)\lambda}\left(\frac{\|\lambda t w^{(t)}\|^2}{t} - u^2\|x_{t+1}\|^2\right)$$

$$= \phi_{t+1}(x_{t+1}^\top w^{(t)}) + \frac{1}{2(t+1)\lambda}\left(2\lambda x_{t+1}^\top w^{(t)} u + \frac{\|\lambda t w^{(t)}\|^2}{t} - u^2\|x_{t+1}\|^2\right)$$

$$= \phi_{t+1}(x_{t+1}^\top w^{(t)}) + \frac{\lambda}{2}\|w^{(t)}\|^2 + \frac{1}{2(t+1)\lambda}\left(2\lambda x_{t+1}^\top w^{(t)} u - \|\lambda w^{(t)}\|^2 - u^2\|x_{t+1}\|^2\right)$$

$$= \phi_{t+1}(w^{(t)\,\top} x_{t+1}) + \frac{\lambda}{2}\|w^{(t)}\|^2 - \frac{\|\lambda w^{(t)} - u x_{t+1}\|^2}{2(t+1)\lambda}\,.$$

Next note that $\|\lambda w^{(t)} - u x_{t+1}\| \leq 2L$ (where we used the triangle inequality, the definition of $w^{(t)}$, and Lemma 3). Therefore,

$$(t+1)D_{t+1}(\alpha^{(t+1)}) - tD_t(\alpha^{(t)}) \geq \phi_{t+1}(w^{(t)\,\top} x_{t+1}) + \frac{\lambda}{2}\|w^{(t)}\|^2 - \frac{2L^2}{(t+1)\lambda}\,.$$

Taking expectation *with respect to the choice of the examples*, and note that the $(t+1)$'th example does not depend on $w^{(t)}$ we obtain that we obtain

$$\mathbb{E}[(t+1)D_{t+1}(\alpha^{(t+1)}) - tD_t(\alpha^{(t)})]$$

$$\geq \mathbb{E}[P_D(w^{(t)})] - \frac{2L^2}{(t+1)\lambda} \geq \mathbb{E}[P_D(w_D^*)] - \frac{2L^2}{(t+1)\lambda}$$

$$\geq \mathbb{E}[P(w^*)] - \frac{2L^2}{(t+1)\lambda} = \mathbb{E}[D(\alpha^*)] - \frac{2L^2}{(t+1)\lambda}.$$

Using Lemma 2 we know that $D_t(\alpha^{(t)}) \ge 0$ for all t. Therefore, by summing the above over t we obtain that

$$\mathbb{E}[nD(\alpha^{(n)})] \ge n\,\mathbb{E}[D(\alpha^*)] - \frac{2L^2\log(en)}{\lambda},$$

which yields

$$\mathbb{E}[D(\alpha^*) - D(\alpha^{(n)})] \le \frac{2L^2\log(en)}{\lambda n}.$$

8.4 Proof of Theorem 4

The proof is identical to the proof of Theorem 1. We just need to notice that at the end of the first stage, we have $\mathbb{E}\epsilon_D^{(0)} \le 2L^2\log(en)/(\lambda n)$. It implies that $t_0 \le \max(0, \lceil n\log(2\lambda n \cdot 2L^2\log(en)/(\lambda nG))\rceil)$.

8.5 Proof of Proposition 1

Consider any feasible dual variable α and the corresponding $w = w(\alpha)$. Since

$$w = \frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i x_i, \qquad w^* = \frac{1}{\lambda n}\sum_{i=1}^{n}\alpha_i^* x_i,$$

we have

$$\lambda(w - w^*)^\top w^* = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \alpha_i^*)w^{*\top}x_i.$$

Therefore

$$
\begin{aligned}
&D(\alpha^*) - D(\alpha)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\phi_i^*(-\alpha_i) - \phi_i^*(-\alpha_i^*) + (\alpha_i - \alpha_i^*)w^{*\top}x_i\right] + \frac{\lambda}{2}[w^\top w - w^{*\top}w^* - 2(w - w^*)^\top w^*]\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\phi_i^*(-\alpha_i) - \phi_i^*(-\alpha_i^*) + (\alpha_i - \alpha_i^*)w^{*\top}x_i\right] + \frac{\lambda}{2}(w - w^*)^\top(w - w^*).
\end{aligned}
$$

Since $w^{*\top}x_i \in \partial\phi_i^*(-\alpha^*)$, we have

$$\phi_i^*(-\alpha_i) - \phi_i^*(-\alpha_i^*) + (\alpha_i - \alpha_i^*)w^{*\top}x_i \ge \gamma_i(\alpha_i - \alpha_i^*)^2.$$

By combining the previous two displayed inequalities, we obtain the first desired bound.

Next, we let $u = w^{*\top}x_i$, $v = w^\top x_i$. Since $-a_i \in \partial\phi_i(v)$ and $-\alpha_i^* \in \partial\phi_i(u)$, it follows that $u \in \partial\phi_i^*(-\alpha_i^*)$ and $v \in \partial\phi_i^*(-a_i)$. Therefore

$$
\begin{aligned}
&|u - v| \cdot |\alpha_i^* - a_i|\\
&= \underbrace{[\phi_i^*(-a_i) - \phi_i^*(-\alpha_i^*) + u(a_i - \alpha_i^*)]}_{\ge 0} + \underbrace{[\phi_i^*(-\alpha_i^*) - \phi_i^*(-a_i) + v(\alpha_i^* - a_i)]}_{\ge 0}\\
&\ge \phi_i^*(-a_i) - \phi_i^*(-\alpha_i^*) + u(a_i - \alpha_i^*) \ge \gamma_i(u)|a_i - \alpha_i^*|^2.
\end{aligned}
$$

This implies the second bound.

8.6 Proof of Theorem 5

The following lemma is very similar to Lemma 1 with nearly identical proof, but it focuses only on the convergence of dual objective function using (5).

Lemma 5. *Assume that* (5) *is valid. Then for any iteration t and any $s \in [0, 1]$ we have*

$$\mathbb{E}[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] \geq \frac{s}{2n}\, \mathbb{E}[D(\alpha^*) - D(\alpha^{(t-1)})] + \frac{3s\lambda}{4n}\|w^* - w^{(t-1)}\|_2^2 - \left(\frac{s}{n}\right)^2 \frac{G_*^{(t)}(s)}{2\lambda} \,,$$

where

$$G_*^{(t)}(s) = \frac{1}{n}\sum_{i=1}^n \left(\|x_i\|_2^2 - \frac{\gamma_i \lambda n}{s}\right) \mathbb{E}[(\alpha_i^* - \alpha_i^{(t-1)})^2].$$

Proof. Since only the i'th element of α is updated, the improvement in the dual objective can be written as

$$n[D(\alpha^{(t)}) - D(\alpha^{(t-1)})] = \underbrace{\left(-\phi^*(-\alpha_i^{(t)}) - \frac{\lambda n}{2}\|w^{(t)}\|^2\right)}_{A_i} - \underbrace{\left(-\phi^*(-\alpha_i^{(t-1)}) - \frac{\lambda n}{2}\|w^{(t-1)}\|^2\right)}_{B_i}.$$

By the definition of the update we have for all $s \in [0, 1]$ that

$$A_i = \max_{\Delta\alpha_i} -\phi^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i\|^2$$

$$\geq -\phi^*(-(\alpha_i^{(t-1)} + s(\alpha_i^* - \alpha_i^{(t-1)}))) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}s(\alpha_i^* - \alpha_i^{(t-1)})x_i\|^2.$$

We can now apply the Jensen's inequality to obtain

$$A_i \geq -s\phi_i^*(-\alpha_i^*) - (1-s)\phi_i^*(-\alpha_i^{(t-1)}) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}s(\alpha_i^* - \alpha_i^{(t-1)})x_i\|^2$$

$$= -s[\phi_i^*(-\alpha_i^*) - \phi_i^*(-\alpha_i^{(t-1)})] \underbrace{-\phi_i^*(-\alpha_i^{(t-1)}) - \frac{\lambda n}{2}\|w^{(t-1)}\|^2}_{B_i} -s(\alpha_i^* - \alpha_i^{(t-1)})x_i^\top w^{(t-1)}$$

$$- \frac{s^2(\alpha_i^* - \alpha_i^{(t-1)})^2}{2\lambda n}\|x_i\|^2.$$

By summing over $i = 1, \ldots, n$, we obtain

$$\sum_{i=1}^n [A_i - B_i] \geq -s\sum_{i=1}^n[\phi_i^*(-\alpha_i^*) - \phi_i^*(-\alpha_i^{(t-1)})] - s\sum_{i=1}^n(\alpha_i^* - \alpha_i^{(t-1)})x_i^\top w^{(t-1)}$$

$$- \frac{s^2}{2\lambda n}\sum_{i=1}^n(\alpha_i^* - \alpha_i^{(t-1)})^2\|x_i\|^2$$

$$= -s\sum_{i=1}^n[\phi_i^*(-\alpha_i^*) - \phi_i^*(-\alpha_i^{(t-1)}) + \lambda(w^* - w^{(t-1)})^\top w^{(t-1)}]$$

$$- \frac{s^2}{2\lambda n}\sum_{i=1}^n(\alpha_i^* - \alpha_i^{(t-1)})^2\|x_i\|^2,$$

where the equality follows from $\sum_{i=1}^{n}(\alpha_i^* - \alpha_i^{(t-1)})x_i = \lambda n(w^* - w^{(t-1)})$. By rearranging the terms on the right hand side using $(w^* - w^{(t-1)})^\top w^{(t-1)} = \|w^*\|_2^2/2 - \|w^{(t-1)}\|_2^2/2 - \|w^* - w^{(t-1)}\|_2^2/2$, we obtain

$$\sum_{i=1}^{n}[A_i - B_i] \geq -s\sum_{i=1}^{n}\left[\phi_i^*(-\alpha_i^*) - \phi_i^*(-\alpha_i^{(t-1)}) + \frac{\lambda}{2}\|w^*\|_2^2 - \frac{\lambda}{2}\|w^{(t-1)}\|_2^2\right] + \frac{\lambda sn}{2}\|w^* - w^{(t-1)}\|_2^2$$

$$- \frac{s^2}{2\lambda n}\sum_{i=1}^{n}(\alpha_i^* - \alpha_i^{(t-1)})^2\|x_i\|^2$$

$$= sn[D(\alpha^*) - D(\alpha^{(t-1)})] + \frac{s\lambda n}{2}\|w^* - w^{(t-1)}\|_2^2 - \frac{s^2}{2\lambda n}\sum_{i=1}^{n}(\alpha_i^* - \alpha_i^{(t-1)})^2\|x_i\|_2^2.$$

We can now apply (5) to obtain

$$\sum_{i=1}^{n}[A_i - B_i] \geq \frac{sn}{2}[D(\alpha^*) - D(\alpha^{(t-1)})] + \frac{3s\lambda n}{4}\|w^* - w^{(t-1)}\|_2^2 - \frac{s^2}{2\lambda n}\sum_{i=1}^{n}(\alpha_i^* - \alpha_i^{(t-1)})^2(\|x_i\|_2^2 - \gamma_i \lambda n/s).$$

This implies the desired result. $\qquad\square$

Lemma 6. *Suppose that for all i, ϕ_i is L-Lipschitz. Let $G_*^{(t)}$ be as defined in Lemma 5. Then*

$$G_*^{(t)}(s) \leq \frac{4L^2 N(s/(\lambda n))}{n}.$$

Proof. Similarly to the proof of Lemma 4, we know that $(\alpha_i^* - \alpha_i^{(t-1)})^2 \leq 4L^2$. Moreover, $\|x_i\|_2^2 \leq 1$, and $\|x_i\|_2^2 - \frac{\gamma_i \lambda n}{s} \leq 0$ when $\gamma_i \geq s/(\lambda n)$. Therefore there are no more than $N(s/(\lambda n))$ data points i such that $\|x_i\|_2^2 - \frac{\gamma_i \lambda n}{s}$ is positive. The desired result follows from these facts. $\qquad\square$

Proof of Theorem 5. Let $\epsilon_D^{(t)} = \mathbb{E}[D(\alpha^*) - D(\alpha^{(t)})]$, and $G_*(s) = 4L^2 N(s/\lambda n)/n$. We obtain from Lemma 5 and Lemma 6 that

$$\epsilon_D^{(t)} \leq (1 - s/(2n))\epsilon_D^{(t-1)} + \left(\frac{s}{n}\right)^2\frac{G_*(s)}{2\lambda}.$$

It follows that for all $t > 0$ we have

$$\epsilon_D^{(t)} \leq (1 - s/(2n))^t\epsilon_D^{(0)} + \frac{1}{1 - (1 - s/(2n))}\left(\frac{s}{n}\right)^2\frac{G_*(s)}{2\lambda}$$

$$\leq e^{-st/2n} + \left(\frac{s}{n}\right)\frac{G_*(s)}{\lambda} \leq e^{-st/2n} + \epsilon_D/2.$$

It follows that when

$$t \geq (2n/s)\log(2/\epsilon_D),$$

we have $\epsilon_D^{(t)} \leq \epsilon_D$. $\qquad\square$

8.7 Proof of Theorem 6

Let $\epsilon_D^{(t)} = \mathbb{E}[D(\alpha^*) - D(\alpha^{(t)})]$. From Proposition 1, we know that for all $t \geq T_0$:

$$\epsilon_D^{(t)} \geq \frac{1}{n} \sum_{i=1}^{n} \left[\gamma_i \, \mathbb{E} \, |\alpha_i^{(t)} - \alpha_i^*|^2 + \frac{\lambda}{2\rho} \, \mathbb{E}((w^{(t)} - w^*)^\top x_i)^2 \right]$$

$$\geq \frac{1}{n} \sum_{i=1}^{n} \left[\gamma_i \, \mathbb{E} \, |\alpha_i^{(t)} - \alpha_i^*|^2 + \frac{\lambda \gamma_i^2}{2\rho} \, \mathbb{E}(u_i^{(t-1)} - \alpha_i^*)^2 \right],$$

where $-u_i^{(t-1)} \in \partial\phi_i(x_i^\top w^{(t)})$. It follows that given any $\gamma > 0$, we have

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \, |\alpha_i^{(t)} - u^{(t-1)}|^2 \leq \frac{N(\gamma)}{n} \sup_i \mathbb{E} \, |\alpha_i^{(t)} - u_i^{(t-1)}|^2 + \frac{2}{n} \sum_{i:\gamma_i \geq \gamma} \left[\mathbb{E} \, |\alpha_i^{(t)} - \alpha_i^*|^2 + \mathbb{E}(u_i^{(t-1)} - \alpha_i^*)^2 \right]$$

$$\leq \frac{N(\gamma)}{n} \sup_i \mathbb{E} \, |\alpha_i^{(t)} - u_i^{(t-1)}|^2 + \frac{\frac{2}{n} \sum_{i=1}^{n} \left[\gamma_i \, \mathbb{E} \, |\alpha_i^{(t)} - \alpha_i^*|^2 + \frac{\lambda \gamma_i^2}{2\rho} \, \mathbb{E}(u_i^{(t-1)} - \alpha_i^*)^2 \right]}{\min(\gamma, \lambda\gamma^2/(2\rho))}$$

$$\leq \frac{N(\gamma)}{n} 4L^2 + \frac{2\epsilon_D^{(t)}}{\min(\gamma, \lambda\gamma^2/(2\rho))},$$

where Lemma 4 is used for the last inequality. Since γ is arbitrary and $\epsilon_D^{(t)} \leq \epsilon_D$, it follows that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \, |\alpha_i^{(t)} - u^{(t-1)}|^2 \leq \tilde{\epsilon}_P.$$

Now plug into Lemma 1, we obtain for all $t \geq T_0 + 1$:

$$\epsilon_D^{(t-1)} - \epsilon_D^{(t)}$$

$$\geq \frac{s}{n} \, \mathbb{E}[P(w^{(t-1)}) - D(\alpha^{(t-1)})] - \left(\frac{s}{n}\right)^2 \frac{1}{2\lambda n} \sum_{i=1}^{n} \mathbb{E}[(u_i^{(t-1)} - \alpha_i^{(t-1)})^2]$$

$$\geq \frac{s}{n} \, \mathbb{E}[P(w^{(t-1)}) - D(\alpha^{(t-1)})] - \left(\frac{s}{n}\right)^2 \frac{\tilde{\epsilon}_P}{2\lambda}.$$

By taking $s = n/T_0$, and summing over $t = T_0 + 1, \ldots, 2T_0 = T$, we obtain

$$\epsilon_D \geq \epsilon_D^{(T_0)} - \epsilon_D^{(T)} \geq \mathbb{E}[P(\bar{w}) - D(\bar{\alpha})] - \frac{\tilde{\epsilon}_P}{2\lambda T_0}.$$

This proves the desired bound.

References

[1] L. Bottou. Stochastic gradient descent examples, Web Page. `http://leon.bottou.org/projects/sgd`.

[2] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. *Journal of Machine Learning Research*, 9:1775–1822, 2008.

[3] C.J. Hsieh, K.W. Chang, C.J. Lin, S.S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *ICML*, pages 408–415, 2008.

[4] D. Hush, P. Kelly, C. Scovel, and I. Steinwart. QP algorithms with guaranteed accuracy and run time for support vector machines. *JMLR*, 7:733–769, 2006.

[5] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[6] T. Joachims. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 216–226, 2006.

[7] Z.Q. Luo and P. Tseng. On the convergence of coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72:7–35, 1992.

[8] O. Mangasarian and D. Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10, 1999.

[9] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *Core discussion papers*, ?(0022010), 2010.

[10] J. C. Platt. Fast training of Support Vector Machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[11] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pages 807–814, 2007.

[12] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l_1 regularized loss minimization. In *ICML*, page 117, 2009.

[13] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.